Journal of Modern Applied Statistical Methods
May, 2004, Vol. 3, No. 1, 27-38

1538 – 9472/04/$95.00

A Power Comparison of Robust Test Statistics Based On Adaptive Estimators

H. J. Keselman
University of Manitoba

Rand R. Wilcox
University of Southern California

James Algina
University of Florida

Katherine Fradette
University of Manitoba

Abdul R. Othman
University of Sains Malysia

Seven test statistics known to be robust to the combined effects of nonnormality and variance heterogeneity were compared for their sensitivity to detect treatment effects in a one-way completely randomized design containing four groups. The six Welch-James-type heteroscedastic tests adopted either symmetric or asymmetric trimmed means, were transformed for skewness, and used a bootstrap method to assess statistical significance. The remaining test, due to Wilcox and Keselman (2003), used a modification of the well-known one-step M-estimator of central tendency rather than trimmed means. The Welch-James-type test is recommended because for nonnormal data likely to be encountered in applied research settings it should be more powerful than the test presented by Wilcox and Keselman. However, the reverse is true for data that are extremely nonnormal.

Key words: Trimmed estimators, symmetric and asymmetric trimming, heterosedastic test statistic, nonnormality, variance heterogeneity

Introduction

Keselman, Wilcox, Othman and Fradette (2002) demonstrated the benefit of testing for symmetry, applying a transformation for skewness, adopting robust estimators and using bootstrapping methodology with a Welch-James-type heteroscedastic statistic in order to obtain a robust test of treatment group equality when data are nonnormal, heterogeneous and unbalanced in one-way completely randomized designs. In particular, they applied a test for symmetry due to Hogg, Fisher and Randles (1975), modified by Babu, Padmanaban and Puri (1999), in order to determine whether data should be trimmed from each tail of the data distribution (symmetric trimming) per group or whether data should only be trimmed from one-tail of the data distribution (asymmetric trimming) per group prior to applying the Johansen (1980) test for treatment group equality. Furthermore, they investigated the utility of transforming the statistic, to circumvent the biasing effects due to skewness, with methods presented by Johnson (1978) and Hall (1992). Lastly, they assessed statistical significance with and without bootstrapping methodology and concluded that critical values obtained through bootstrapping provided an additional benefit against the deleterious effects of nonnormality and variance heterogeneity.

These authors concluded by recommending that researchers test for treatment group equality by adopting the aforementioned

H. J. Keselman (kesel@ms.umanitoba.ca) is Professor of Psychology. Rand R. Wilcox (rwilcox@usc.edu) is Professor of Psychology. James Algina (algina@ufl.edu) is Professor of Educational Psychology. Katherine Fradette (umfradet@cc.umanitoba.ca) is a graduate student in psychology. Her interests are in applied statistical analysis. Abdul Rahman Othman (oarahman@usm.my) is Associate Professor, School of Distance Education. Work on this project was supported by a grant by the Natural Sciences and Engineering Council of Canada.

modifications to the Johansen test with 10% symmetric trimming or 20% asymmetric trimming based on a preliminary test for symmetry. They noted as well that other percentages of symmetric/asymmetric trimming worked quite well with respect to Type I error control (e.g., 15%/30%).

Othman, Keselman, Padmanabhan, Wilcox, and Fradette (2003) compared a number of recently developed adaptive robust methods with respect to their ability to control Type I errors and their sensitivity to detect differences between groups when data were nonnormal, heterogeneous, and the design was unbalanced. In particular, two new approaches to comparing the typical score across treatment groups due to Babu et al. (1999) were compared to two new methods presented by Wilcox and Keselman (2003) and Keselman et al. (2002). The procedures examined exhibited very good Type I error control and the power results clearly favored one of the methods (a method they referred to as MOMT) presented by Wilcox and Keselman; indeed, in the vast majority of the cases investigated, this most favored approach had substantially larger power values compared to the other procedures.

Based on the findings of these two studies an important research question remains. Namely, how does the power of the robust and powerful procedure investigated by Othman et al. (2003) (i.e., MOMT) compare to the sensitivity of the Johansen (1980) Welch-James-(WJ)-type procedure for detecting treatment effects in one-way completely randomized designs? This question is important because other investigators have recommended the WJ test due to its sensitivity to detect effects for other designs [See e.g., Algina & Keselman (1998)] and neither Keselman et al. (2002) or Othman et al. investigated the power of the WJ test.

Test Statistics

The WJ Statistic

Lix and Keselman (1995) showed how the various Welch (1938, 1951) statistics that appear in the literature for testing omnibus main and interaction effects as well as focused hypotheses using contrasts in univariate and multivariate independent and correlated groups designs can be formulated from a general linear model perspective, thus allowing researchers to apply one statistical procedure to any testable model effect. Their approach is adopted in this article and is presented in abbreviated form.

A general approach for testing hypotheses of mean equality using an approximate degrees of freedom solution is developed using matrix notation. The multivariate perspective is considered first; the univariate model is a special case of the multivariate. Consider the general linear model:

$$Y = X\beta + \xi \ , \qquad (1)$$

where Y is an $N \times p$ matrix of scores on p dependent variables or p repeated measurements, N is the total sample size, X is an $N \times r$ design matrix consisting entirely of zeros and ones with $\text{rank}(X) = r$, β is an $r \times p$ matrix of nonrandom parameters (i.e., population means), and ξ is an $N \times p$ matrix of random error components. Let Y_j $(j = 1,\ldots,r)$ denote the submatrix of Y containing the scores associated with the n subjects in the jth group (cell) (For the one-way design considered in this paper $n = n_j$). It is typically assumed that the rows of Y are independently and normally distributed, with mean vector β_j and variance-covariance matrix Σ_j [i.e., $N(\beta_j, \Sigma_j)$], where the jth row of β, $\beta_j = [\mu_{j1} \ldots \mu_{jp}]$, and $\Sigma_j \neq \Sigma_{j'} (j \neq j')$. Specific formulas for estimating β and Σ_j, as well as an elaboration of Y are given in Lix and Keselman (1995, See their Appendix A).

The general linear hypothesis is

$$H_0 : R\mu = 0 \ , \qquad (2)$$

where $R = C \otimes U^T$, C is a $df_C \times r$ matrix which controls contrasts on the independent groups effect(s), with $\text{rank}(C) = df_C \leq r$, and U is a $p \times df_U$ matrix which controls contrasts on the within-subjects effect(s), with rank $(U) = df_U \leq p$, '$\otimes$' is the Kronecker or direct

product function, and '^{T}' is the transpose operator. For multivariate independent groups designs, U is an identity matrix of dimension p (i.e., I_p). The R contrast matrix has $df_C \times df_U$ rows and $r \times p$ columns. In Equation 2, $\mu = vec(\beta^T) = [\beta_1 \ldots \beta_r]^T$. In other words, μ is the column vector with $r \times p$ elements obtained by stacking the columns of β^T. The 0 column vector is of order $df_C \times df_U$ [See Lix & Keselman (1995) for illustrative examples].

The generalized test statistic given by Johansen (1980) is

$$T_{WJ} = (R\hat{\mu})^T (R\hat{\Sigma}R^T)^{-1}(R\hat{\mu}) \qquad (3)$$

where $\hat{\mu}$ estimates μ, and $\hat{\Sigma} = diag[\hat{\Sigma}_1 / n_1 \ldots \hat{\Sigma}_r / n_r]$, a block matrix with diagonal elements $\hat{\Sigma}_j / n_j$.

This statistic, divided by a constant, c (i.e., T_{WJ} / c), approximately follows an F distribution with degrees of freedom $\nu_1 = df_C \times df_U$, and $\nu_2 = \nu_1(\nu_1 + 2)/(3A)$, where $c = \nu_1 + 2A - (6A)/(\nu_1 + 2)$. The formula for the statistic A is provided in Lix and Keselman (1995).

When $p = 1$, that is, for a univariate model, the elements of Y are assumed to be independently and normally distributed with mean μ_j and variance σ_j^2 [i.e., $N(\mu_j, \sigma_j^2)$]. To test the general linear hypothesis, C has the same form and function as for the multivariate case, but now $U = 1, \hat{\mu} = [\hat{\mu}_1 \ldots \hat{\mu}_r]^T$ and $\hat{\Sigma} = diag[\hat{\sigma}_1^2 / n_1 \ldots \hat{\sigma}_r^2 / n_r]$, (See Lix & Keselman's 1995 Appendix A for further details of the univariate model.).

Robust Estimation

In this article robust estimates of central tendency and variability are applied to the T_{WJ} statistic. That is, heteroscedastic ANOVA methods are readily extended to the problem of comparing trimmed means. The goal is to determine whether the effect of a treatment varies across J $(j = 1, \ldots, J)$ groups; that is, to determine whether a typical score varies across groups. When trimmed means are being compared the null hypothesis pertains to the equality of population trimmed means, i.e., the $\mu_t s$. That is, to test the omnibus hypothesis in a one-way completely randomized design, the null hypothesis would be $H_0 : \mu_{t1} = \mu_{t2} = \cdots = \mu_{tJ}$.

Let $Y_{(1)j} \le Y_{(2)j} \le \cdots \le Y_{(n_j)j}$ represent the ordered observations associated with the jth group. Let $g_j = [\gamma n_j]$, where γ represents the proportion of observations that are to be trimmed in each tail of the distribution and $[x]$ is the greatest integer $\le x$. The effective sample size for the jth group becomes $h_j = n_j - 2g$. The jth sample trimmed mean is

$$\hat{\mu}_{tj} = \frac{1}{h_j} \sum_{i=g_j+1}^{n_j - g_j} Y_{(i)j} \ . \qquad (4)$$

Wilcox (1995) suggested that 20% trimming should be used. (See Wilcox, 1995, and the references cited for a justification of the 20% rule.)

The sample Winsorized mean is necessary and is computed as

$$\hat{\mu}_{wj} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij} \ , \qquad (5)$$

where

$$\begin{aligned} X_{ij} &= Y_{(g_j+1)j} \ \textit{if} \ Y_{ij} \le Y_{(g_j+1)j} \\ &= Y_{ij} \ \text{if} \ Y_{(g_j+1)j} < Y_{ij} < Y_{(n_j - g_j)j} \\ &= Y_{(n_j - g_j)j} \ \text{if} \ Y_{ij} \ge Y_{(n_j - g_j)j} \ . \end{aligned}$$

The sample Winsorized variance, which is required to get a theoretically valid estimate of the standard error of a trimmed mean, is then given by

$$\hat{\sigma}_{wj}^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_{ij} - \hat{\mu}_{wj})^2 \ . \qquad (6)$$

The standard error of the trimmed mean is estimated with $\sqrt{(n_j-1)\hat{\sigma}_{wj}^2/[h_j(h_j-1)]}$.

Under asymmetric trimming, and assuming, without loss of generality, that the distribution is positively skewed so that trimming takes place in the upper tail, the *j*th sample trimmed mean is

$$\hat{\mu}_{tj}=\frac{1}{h_j}\sum_{i=1}^{n_j-g_j}Y_{(i)j}$$

and the *j*th sample Winsorized mean is

$$\hat{\mu}_{wj}=\frac{1}{n_j}\sum_{i=1}^{n_j}X_{ij}\ ,$$

where

$$X_{ij}=Y_{ij}\ \textit{if}\ Y_{ij}\le Y_{(n_j-g_j)j}$$
$$=Y_{(n_j-g_j)j}\ \text{if}\ Y_{ij}\ge Y_{(n_j-g_j)j}\ .$$

The sample Winsorized variance is again defined as (given the new definition of $\hat{\mu}_{wj}$)

$$\hat{\sigma}_{wj}^2=\frac{1}{n_j-1}\sum_{i=1}^{n_j}(X_{ij}-\hat{\mu}_{wj})^2$$

and the standard error of the mean again takes its usual form (given the new definition of $\hat{\mu}_{wj}$).

Thus, with robust estimation, the trimmed group means ($\hat{\mu}_{tj}s$) replace the least squares group means ($\hat{\mu}_j s$), the Winsorized group variances estimators ($\hat{\sigma}_{wj}^2 s$) replace the least squares variances ($\hat{\sigma}_j^2 s$), and h_j replaces n_j and accordingly one computes the robust version of T_{WJ}, T_{WJt} (See Keselman, Wilcox, & Lix, 2003; and Rocke, Downs & Rocke, 1982, for another justification for adopting robust estimates).

Bootstrapping

Now considered is how extensions of the ANOVA method just outlined might be improved. In terms of probability coverage and controlling the probability of a Type I error, extant investigations indicate that the most successful method, when using a 20% trimmed mean (or some M-estimator), is some type of bootstrap method.

Following Westfall and Young (1993), and as described by Wilcox (1997), let $C_{ij}=Y_{ij}-\hat{\mu}_{tj}$; thus, the C_{ij} values are the empirical distribution of the *j*th group, centered so that the sample trimmed mean is zero. That is, *the empirical distributions are shifted so that the null hypothesis of equal trimmed means is true in the sample*. The strategy behind the bootstrap is to use the shifted empirical distributions to estimate an appropriate critical value. For each *j*, obtain a bootstrap sample by randomly sampling with replacement n_j observations from the C_{ij} values, yielding $Y_1^*,\ldots,Y_{n_j}^*$. Let T_{WJt}^* be the value of Johansen's (1980) test based on the bootstrap sample. Now randomly sample (with replacement), B bootstrap samples from the shifted/centered distributions each time calculating the statistic T_{WJt}^*. The B values of T_{WJt}^* are put in ascending order, that is, $T_{WJt(1)}^*\le\cdots\le T_{WJt(B)}^*$, and an estimate of an appropriate critical value is $T_{WJt(a)}^*$, where $a=(1-\alpha)B$, rounded to the nearest integer. One will reject the null hypothesis of location equality (i.e., $H_0:\mu_{t1}=\mu_{t2}=\cdots=\mu_{tJ}$) when $T_{WJt}>T_{WJt(a)}^*$, where T_{WJt} is the value of the heteroscedastic statistic based on the original non-bootstrapped data. Keselman et al. (2002) illustrate the use of this procedure for testing both omnibus and sub-effect (linear contrast) hypotheses in completely randomized and correlated groups designs.

Transformations for the Welch-James Statistic

Guo and Luh (2000) and Luh and Guo 1999 found that Johnson's (1978) and Hall's (1992) transformations improved the performance of several heteroscedastic test statistics when they were used with trimmed means, including the WJ statistic, in the presence of heavy-tailed and skewed distributions.

In this study both approaches are compared for removing skewness when applied

to the T_{WJt} statistic. Let $Y_{ij} = (Y_{1j}, Y_{2j}, \ldots, Y_{n_j j})$ be a random sample from the *j*th distribution. Let $\hat{\mu}_{tj}$, $\hat{\mu}_{wj}$ and $\hat{\sigma}^2_{wj}$ be, respectively, the trimmed mean, Winsorized mean and Winsorized variance of group *j*. Define the Winsorized third central moment of group *j* as

$$\hat{\mu}_{3j} = \frac{1}{n_j}\sum_{i=1}^{n_j}(X_{ij} - \hat{\mu}_{wj})^3 \ .$$

Let

$$\tilde{\sigma}^2_{wj} = \frac{(n_j - 1)}{h_j - 1}\hat{\sigma}^2_{wj} \ ,$$

$$\tilde{\mu}_{wj} = \frac{n_j}{h_j}\hat{\mu}_{3j} \ ,$$

$$q_j = \frac{\tilde{\sigma}^2_{wj}}{h_j} \ ,$$

$$w_{tj} = \frac{1}{q_j} \ ,$$

$$U_t = \sum_{j=1}^{J} w_{tj} \ ,$$

and

$$\hat{\mu}_t = \frac{1}{U_t}\sum_{j=1}^{J} w_{tj}\hat{\mu}_{tj} \ .$$

Luh and Guo (2000) defined a trimmed mean statistic with Johnson's transformation as

$$T_{Johnson_j} = (\hat{\mu}_{tj} - \hat{\mu}_t) + \frac{\tilde{\mu}_{wj}}{6\tilde{\sigma}^2_{wj}h_j} + \frac{\tilde{\mu}_{wj}}{3\tilde{\sigma}^4_{wj}}(\hat{\mu}_{tj} - \hat{\mu}_t)^2. \qquad (7)$$

From Guo and Luh (2000) one can deduce that a trimmed mean statistic with Hall's (1992) transformation would be

$$T_{Hall_j} = (\hat{\mu}_{tj} - \hat{\mu}_t) + \frac{\tilde{\mu}_{wj}}{6\tilde{\sigma}^2_{wj}h_j} + \frac{\tilde{\mu}_{wj}}{3\tilde{\sigma}^4_{wj}}(\hat{\mu}_{tj} - \hat{\mu}_t)^2 + \frac{\tilde{\mu}^2_{wj}}{27\tilde{\sigma}^8_{wj}}(\hat{\mu}_{tj} - \hat{\mu}_t)^3 \ . \ (8)$$

Keselman et al. (2002) indicated that sample trimmed means, sample Winsorized variances and trimmed sample sizes can be substituted for the usual sample means, variances and sample sizes in the T_{WJ} statistic. That is,

$$T_{WJ} = \sum_{j=1}^{J} w_{tj}(\hat{\mu}_{tj} - \hat{\mu}_t)^2 \ ,$$

which, when divided by *c*, is distributed as an *F* variable with df of $J - 1$ and

$$\nu = (J^2 - 1)\left[3\sum_{j=1}^{J}\frac{(1 - w_{tj}/U_t)^2}{h_j - 1}\right]^{-1}$$

where

$$c = (J - 1)\left[1 + \frac{2(J-2)}{J^2 - 1}\sum_{j=1}^{J}\frac{(1 - w_{tj}/U_t)^2}{h_j - 1}\right] .$$

Now we can define

$$T_{WJ_{Johnson}} = \sum_{j=1}^{J} w_{tj}(T_{Johnson_j})^2 \ , \qquad (9)$$

and

$$T_{WJ_{Hall}} = \sum_{j=1}^{J} w_{tj}(T_{Hall_j})^2 \ . \qquad (10)$$

Then $T_{WJ_{Johnson}}$ and $T_{WJ_{Hall}}$, when divided by *c*, are also distributed as *F* variates with no change in degrees of freedom.

A Preliminary Test for Symmetry

A stumbling block to adopting asymmetric versus symmetric trimming has been the inability of researchers to determine when to adopt one form of trimming over the other. Work by Hogg et al. (1975) and Babu et al. (1999), however, may provide a successful solution to this problem. The details of this method are presented in Othman et al. (2002).

The One-Step Modified M-(MOM) Estimator

For *J* independent groups (this estimator can also be applied to dependent groups) consider the MOM estimator introduced by Wilcox and Keselman (2003). They suggested modifying the well-known one-step M-estimator

$$\frac{1.28(MADN_j)(i_2 - i_1) + \sum_{i=i_1+1}^{n_j - i_2} Y_{(i)j}}{n_j - i_1 - i_2}, \quad (11)$$

by removing $1.28\,(MADN_j)(i_2 - i_1)$, where $MADN_j = MAD_j/.6745$, $MAD_j =$ the median of the values $|Y_{ij} - \hat{M}_j|,\ldots,|Y_{n_j j} - \hat{M}_j|$, $\hat{M}_j$ is the median of the jth group, $i_1 =$ the number of observations where $Y_{ij} - \hat{M}_j < 2.24(MADN_j)$ and $i_2 =$ the number of observations where $Y_{ij} - \hat{M}_j > 2.24(MADN_j)$. Thus, the modified M-estimator suggested by Wilcox and Keselman is

$$\hat{\theta}_j = \sum_{i=i_1+1}^{n_j - i_2} \frac{Y_{(i)j}}{n_j - i_1 - i_2}. \quad (12)$$

The MOM estimate of location is just the average of the values left after all outliers (if any) are discarded. The constant 2.24 is motivated in part by the goal of having a reasonably small standard error when sampling from a normal distribution. Moreover, detecting outliers with Equation 12 is a special case of a more general outlier detection method derived by Rousseeuw and van Zomeren (1990).

MOMT

MOM estimators, like trimmed means, can be applied to test statistics to investigate the equality of this measure (θ) of the typical score across treatment groups. The null hypothesis is

$$H_0: \theta_1 = \theta_2 = \cdots = \theta_J, \quad (13)$$

where θ_j is the population value of MOM associated with the jth group. Of the two statistics that can be used to test this hypothesis, Othman et al. (in press) found that the one based on the work of Liu and Singh (1997) was most powerful. To obtain the test, let

$$\delta_{jj'} = \theta_j - \theta_{j'} \ (j < j'). \quad (14)$$

Thus, the $\delta_{jj'}$s are the all possible pairwise comparisons among the J treatment groups. Now, if all groups have a common measure of location (i.e., $\theta_1 = \theta_2 = \cdots = \theta_J$), then $H_0: \delta_{12} = \delta_{13} = \cdots = \delta_{J-1,J} = 0$. A bootstrap method can be used to assess statistical significance. Bootstrap samples are obtained for the Y_{ij} values and one rejects if the zero vector is sufficiently far from the center of the bootstrap estimates of the delta values. Thus, bootstrap samples are obtained from the Y_{ij} values rather than the C_{ij}s. For each bootstrap replication ($B = 599$ is recommended) one computes the robust estimators (i.e., MOM) of location (i.e., $\hat{\theta}^*_{jb}, j = 1,\ldots,J;\ b = 1,\ldots,B$) and the corresponding estimates of $\delta_{jj'b}$ ($\hat{\delta}^*_{jj'b} = \hat{\theta}^*_{jb} - \hat{\theta}^*_{j'b}$). The strategy is to determine how deeply $0 = (0\ 0\ \ldots\ 0)$ is nested within the bootstrap values $\hat{\delta}^*_{jj'b}$, where 0 is a vector having length $K = J(J-1)/2$. This assessment is made by adopting a modification of Mahalanobis's distance statistic.

For notational convenience, the K differences $\hat{\delta}_{jj'}$ can be rewritten as $\hat{\Delta}_1,\ldots,\hat{\Delta}_K$ and their corresponding bootstrap values as $\hat{\Delta}^*_{kb}$ ($k = 1,\ldots,K;\ b = 1,\ldots,B$). Thus, let

$$\bar{\Delta}^*_k = \frac{1}{B}\sum_{b=1}^{B} \hat{\Delta}^*_{kb},$$

and

$$Z_{kb} = \hat{\Delta}^*_{kb} - \bar{\Delta}^*_k + \hat{\Delta}_k.$$

(Note the Z_{kb}s are shifted bootstrap values having mean $\hat{\Delta}_k$.) Now define

$$S_{kk'} = \frac{1}{B-1}\sum_{b=1}^{B}(Z_{kb} - \bar{Z}_k)(Z_{k'b} - \bar{Z}_{k'}), \quad (15)$$

where

$$\bar{Z}_k = \frac{1}{B}\sum_{b=1}^{B} Z_{kb}.$$

(Note: The bootstrap population mean of $\bar{\Delta}^*_k$ is known and is equal to $\hat{\Delta}_k$.)

With this procedure, next compute

$$D_b = (\hat{\Delta}_b^* - \hat{\Delta})S^{-1}(\hat{\Delta}_b^* - \hat{\Delta})' \quad , \qquad (16)$$

where $\hat{\Delta}_b^* = (\hat{\Delta}_{1b}^*, \ldots, \hat{\Delta}_{Kb}^*)$ and $\hat{\Delta} = (\hat{\Delta}_1, \ldots, \hat{\Delta}_K)$. Accordingly, D_b measures how closely $\hat{\Delta}_b$ is located to $\hat{\Delta}$. If the null vector (0) is relatively far from $\hat{\Delta}$ one rejects H_0. Therefore, to assess statistical significance, put the D_b values in ascending order $(D_{(1)} \leq \cdots \leq D_{(B)})$ and let $a = (1-\alpha)B$ (rounded to the nearest integer). Reject H_0 if

$$T \geq D_{(a)} \quad , \qquad (17)$$

where

$$T = (0 - \hat{\Delta})S^{-1}(0 - \hat{\Delta})' \quad . \qquad (18)$$

It is important to note that $\theta_1 = \theta_2 = \cdots = \theta_J$ can be true *iff* $H_0 : \theta_1 - \theta_2 = \cdots = \theta_{J-1} - \theta_J = 0$ (Therefore, it suffices to test that a set of *K* pairwise differences equal zero.) However, to avoid the problem of arriving at different conclusions (i.e., sensitivity to detect effects) based on how groups are arranged (if all MOMs are unequal), it is recommended that one test the hypothesis that all pairwise differences equal zero.

Methodology

Seven tests for treatment group equality were compared for their sensitivity to detect treatment effects under conditions of nonnormality and variance heterogeneity in an independent groups design with four treatments. The procedures investigated, based on the findings and recommendations of Keselman et al. (2002) and Othman et al. (in press), were:

WJ with preliminary testing for symmetry (Babu et al., 1999)/Symmetric and Asymmetric Trimming:

1.-3. WJJB1020(1530)(2040)-WJ with Johnson's (1978) transformation and bootstrapping. If data are symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 20% (30%) (40%) one sided trimming.

4.-6. WJHB1020(1530)(2040)-WJ with Hall's (1990) transformation and bootstrapping. If data are symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 20% (30%) (40%) one sided trimming.

7. MOMT.

Four variables were manipulated in the study: (a) sample size, (b) degree of variance heterogeneity, (c) pairing of unequal variances and group sizes, and (d) population distribution.

An unbalanced completely randomized design containing four groups was investigated since previous research has looked at this design (e.g., Keselman et al., 2002; Lix & Keselman, 1998; Othman et al., in press; Wilcox, 1988). The two cases of total sample size and the group sizes were $N = 70$ (10, 15, 20, 25) and $N = 90$ (15, 20, 25, 30). The values of n_j were selected from those used by Lix and Keselman (1998) in their study comparing omnibus tests for treatment group equality; their choice of values was, in part, based on having group sizes that others have found to be generally sufficient to provide reasonably effective Type I error control (e.g., see Wilcox, 1994).

The unequal variances were either in a 36:1:1:1 or 8:1:1:1 ratio. Though a ratio of 36:1:1:1 may seem extreme, ratios similar to this case, and larger, have been reported in the literature. Keselman, et al. (1998) after reviewing articles published in prominent education and psychology journals noted that they found ratios as large as 24:1 and 29:1 in one-way and factorial completely randomized designs. Wilcox (2003) cited data sets where the ratio was 17,977:1!

It is appropriate to compare the test statistics under this condition of variance heterogeneity -- the results under this condition will tell how the tests perform under conditions that either have been reported or may likely be encountered with actual data sets. Furthermore, even assuming that a 36:1:1:1 ratio of variances may be large, it nonetheless seems reasonable to see how well the tests perform under a potentially extreme condition. This will provide researchers with information regarding how well the tests hold up under any degree of heterogeneity they are likely to obtain in their

data, thus providing a generalizable result. Nonetheless, the tests were also compared under a less extreme condition of heterogeneity, i. e., when the variances were in a ratio of 8:1:1:1.

Variances and group sizes were both positively and negatively paired. A positive pairing referred to the case in which the largest n_j was associated with the population having the largest variance; a negative pairing referred to the case in which the largest n_j was associated with the population having the smallest variance. These conditions were chosen since they typically produce conservative and liberal results, respectively.

With respect to the effects of distributional shape on Type I error, we chose to investigate nonnormal distributions in which the data were obtained from a variety of skewed distributions. In addition to generating data from a χ^2_3 distribution, we also used the method described in Hoaglin (1985) to generate distributions with more extreme degrees of skewness and kurtosis. These particular types of nonnormal distributions were selected since educational and psychological research data typically have skewed distributions (Micceri, 1989; Wilcox, 1994). Furthermore, Sawilowsky and Blair (1992) investigated the effects of eight nonnormal distributions, which were identified by Micceri on the robustness of Student's *t* test, and they found that only distributions with the most extreme degree of skewness (e.g., $\gamma = 1.64$) affected the Type I error control of the independent sample *t* statistic. Thus, because the statistics investigated have operating characteristics similar to those reported for the *t* statistic, it was assumed that this approach to modeling skewed data would adequately reflect conditions in which those statistics might not perform optimally.

For the χ^2_3 distribution, skewness and kurtosis values are $\gamma_1 = 1.63$ and $\gamma_2 = 4.00$, respectively. The other nonnormal distributions were generated from the *g* and *h* distribution (Hoaglin, 1985). Specifically, two *g* and *h* distributions were investigated: (a) $g = .5$ and $h = 0$ and (b) $g = .5$ and $h = .5$, where *g* and *h* are parameters that determine the moments of a distribution. To give meaning to these values it should be noted that for the standard normal distribution $g = h = 0$. Thus, when $g = 0$ a distribution is symmetric, and the tails of a distribution will become heavier as *h* increases in value. Values of skewness and kurtosis corresponding to the investigated values of *g* and *h* are (a) $\gamma_1 = 1.75$ and $\gamma_2 = 8.9$, respectively, and (b) $\gamma_1 = \gamma_2 =$ undefined.

These values of skewness and kurtosis for the *g* and *h* distributions are theoretical values; Wilcox (1997, p. 73) reported computer generated values, based on 100,000 observations; $\hat{\gamma}_1 = 1.81$ and $\hat{\gamma}_2 = 9.7$ for $g = .5$ and $h = 0$ and $\hat{\gamma}_1 = 120.10$ and $\hat{\gamma}_2 = 18{,}393.6$ for $g = .5$ and $h = .5$. Thus, the conditions investigated could be described as extreme. They are intended to indicate the operating characteristics of the procedures under substantial departures from homogeneity and normality, with the premise that, if a procedure works under the most extreme of conditions, it is likely to work under most conditions likely to be encountered by researchers.

In terms of the data generation procedure, to obtain pseudo-random normal variates, the SAS generator RANNOR (SAS Institute, 1989) was used. If Z_{ij} is a standard unit normal variate, then $Y_{ij} = \mu_j + \sigma_j \times Z_{ij}$ is a normal variate with mean equal to μ_j and variance equal to σ_j^2. To generate pseudo-random variates having a χ^2 distribution with three degrees of freedom, three standard normal variates were squared and summed.

To generate data from a *g*- and *h*-distribution, standard unit normal variables were converted to random variables via

$$Y_{ij} = \frac{\exp(gZ_{ij}) - 1}{g} \exp\left(\frac{hZ_{ij}^2}{2}\right),$$

according to the values of *g* and *h* selected for investigation. To obtain a distribution with standard deviation σ_j, each Y_{ij} was multiplied by a value of σ_j. It is important to note that this

does not affect the value of the mean when $g = 0$ (see Wilcox, 1994, p. 297). However, when $g > 0$, the population mean for a *g*- and *h*-distributed variable is

$$\mu_{gh} = \frac{1}{g(1-h)^{1/2}}(e^{g^2/2(1-h)} - 1)$$

(see Hoaglin, 1985, p. 503). Thus, for those conditions where $g > 0$, μ_{tj} was first subtracted from Y_{ij} before multiplying by σ_j. When working with MOMs, θ_j was first subtracted from each observation (The value of θ_j was obtained from generated data from the respective distributions based on one million observations.). Specifically, for procedures using trimmed means, μ_{tj} was subtracted from the generated variates under every generated distribution. Correspondingly, for the procedure based on MOMs, θ_j was subtracted for all distributions investigated.

The standard deviation of a *g*- and *h*-distribution is not equal to one, and thus the values reflect only the amount by which each random variable is multiplied and not the actual values of the standard deviations (see Wilcox, 1994, p. 298). As Wilcox noted, the values for the variances (standard deviations) more aptly reflect the ratio of the variances (standard deviations) between the groups. Five thousand replications of each condition were performed using a .05 statistical significance level. According to Wilcox (1997) and Hall (1986), *B* was set at 599; that is, their results suggest that it may be advantageous to choose *B* such that $1-\alpha$ is a multiple of $(B+1)^{-1}$.

Lastly, the power of the tests were compared by selected constants to be added to the observations in each group, to avoid ceiling and floor effects; however, values were also selected based on the work of Cohen (1988, pp. 270-272). Specifically, a range for the difference between the groups was selected and then specified this range according to a minimum-, equal-, or maximum-variability difference between the groups. Accordingly, the constants that were added (after centering the data) to the randomly generated data in the four groups were $-1, 0, 0, 1$ (minimum variability), $-1, -.5, .5, 1$ (equal variability), and $-1, -1, 1, 1$ (maximum variability).

Results

Prior to the presentation of power results, the reader should be reminded that the tests examined, very effectively control Type I errors under the conditions studied in this investigation; the Type I error results have been reported in Keselman et al. (2002) and Othman et al. (in press).

The preliminary analysis of the empirical power rates indicated that there were only relatively minor differences between the WJ tests due to type of transformation [i.e., Johnson (1978) or Hall (1992)] for skewness. Accordingly, in Table 1, which contains the empirical power rates, the values tabled for the WJ procedure are based on averaging over the two WJ tests employing the two different transformations for skewness.

Furthermore, no differences existed between the procedures due to sample size and accordingly, the tabled values have been averaged over the two cases of sample size for each test investigated. As well, we note that power rates have been averaged over the type of range investigated (i.e., minimum-, equal- and maximum-variability). Researchers certainly would not be privy to this type of information and thus it seems most reasonable to collapse over this variable.

Based on the values contained in Table 1 we note that: (1) either the WJ1530 and/or the WJ2040 procedure was always at least as powerful as the WH1020 test, (2) the WJ2040 test was at least as powerful as the WJ1530 test for two of the nonnormal distributions investigated (χ^2_3 and $g = .5$ and $h = 0$), while it was marginally less powerful for the remaining nonnormal distribution investigated ($g = .5$ and $h = .5$), and (3) the WJ tests were more powerful than the MOMT test for the χ^2_3 and $g = .5$ and $h = 0$ nonnormal distributions, yet less powerful when the data were $g = .5$ and $h = .5$ distributed.

Table 1. Power Values

Distribution	Max σ^2	Pairing	WJ1020	WJ1530	WJ2040	MOMT	WJ1530-MOMT	WJ2040-MOMT
Chi-Squared	8	Pos	57	60	65	38	22	27
Chi-Squared	36	Pos	52	55	59	34	21	25
Chi-Squared	8	Neg	54	56	61	42	14	19
Chi-Squared	36	Neg	49	50	54	38	12	16
g=.5/h=0	8	Pos	93	94	94	87	07	07
g=.5/h=0	36	Pos	88	90	90	81	09	09
g=.5/h=0	8	Neg	95	95	93	92	03	01
g=.5/h=0	36	Neg	92	92	89	89	03	0
g=.5/h=.5	8	Pos	68	71	69	76	-05	-07
g=.5/h=.5	36	Pos	62	65	64	68	-03	-04
g=.5/h=.5	8	Neg	68	71	68	81	-10	-13
g=.5/h=.5	36	Neg	63	67	65	76	-09	-11
Average			70	72	73	67		

The table also includes values indicating the difference in powers between the WJ1530 and WJ2040 tests and the MOMT test (notated as WJ1530-MOMT and WJ2040 – MOMT). These difference scores indicate that power differences favoring the WJ tests were as large as 27 percentage points while those favoring MOMT were at times more powerful by 13 percentage points.

Conclusion

Keselman et al (2002) noted that researchers could achieve robustness to nonnormality and variance heterogeneity by using trimmed means in a heteroscedastic test statistic [i.e., Johansen (1980)] when data were either trimmed symmetrically or asymmetrically based on a preliminary test for symmetry due to Hogg et al. (1975) and Babu et al. (1999) and when the test was modified by a transformation for skewness due either to Johnson (1978) or Hall (1992) and when statistical significance was assessed through a bootstrap method.

Othman et al. (in press) found that when treatment group equality was assessed with a test statistic suggested by Liu and Singh (1997) comparing across groups a measure of central tendency based on Wilcox and Keselman's (2003) modification of the well-known one-step

M-estimator (i.e., MOM), Type I errors were very effectively controlled under very adverse conditions of nonnormality and variance heterogeneity. Furthermore, and most important to the motivation for the current investigation, they also found that the procedure was substantially more powerful than the other test statistics they investigated.

The purpose of this investigation therefore was to contrast the sensitivity of the test examined by Othman et al. (in press) with the Johansen (1980) Welch-James-type procedure investigated by Keselman et al. (2002) since both methods provide very good Type I error control and good power characteristics have been attributed to the WJ-type test by other researchers (see e.g., Algina & Keselman, 1998), though it has not been compared to the MOMT test nor under conditions examined by Keselman et al. and Othman et al.

For the three nonnormal distributions investigated, it was found that the WJ-type tests were more powerful than the MOMT test when data were moderately to substantially nonnormal (i.e., χ^2_3 and $g = .5$ and $h = 0$ distributed); however, when the data were extremely nonnormal (i.e., $g = .5$ and $h = .5$ distributed), the MOMT test was more powerful than the WJ-type tests. In the former case, the differences favored the WJ-type tests by as much as 27 percentage points while in the latter case MOMT values, at times, exceeded the WJ values by as much as 13 percentage points.

Based on these findings, we recommend, in general, the WJ-type tests that utilize symmetric or asymmetric trimmed means (with the type of trimming based on the Babu et al., 1999, test for symmetry) with a transformation for skewness (due either to Johnson, 1978, or Hall, 1992) and where statistical significance is assessed through the bootstrap method defined in this article (or in Keselman et al., 2002). In particular, the WJ2040 method is recommended. That is, for most nonnormal distributions that researchers are likely to encounter in applied work it is not likely that their data will be as nonnormal as that characterized by the $g = .5$ and $h = .5$ distribution, and thus they are likely to have greater sensitivity to detect treatment effects with the WJ-type test than with the MOMT test. However, when researchers suspect that their data is extremely nonnormal, in a manner similar to the characteristics of the $g = .5$ and $h = .5$ distribution, then clearly, it will be advantageous to adopt the MOMT test. Numerical results for MOMT can be obtained from Wilcox (2003, pp. 84, 314).

References

Algina, J., & Keselman, H. J. (1998). A power comparison of the Welch-James and Improved General Approximation tests in the split-plot design. *Journal of Educational and Behavioral Statistics*, *23*, 152-169.

Babu, J. G., Padmanabhan A. R., & Puri, M. P. (1999). Robust one-way ANOVA under possibly non-regular conditions. *Biometrical Journal*, *413*, 321-339.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Guo, J. H., & Luh, W. M. (2000). An invertible transformation two-sample trimmed t-statistic under heterogeneity and nonnormality. *Statistics & Probability Letters*, *49*, 1-7.

Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics*, *14*, 1431-1452.

Hall, P. (1992). On the removal of skewness by transformation. *Journal of the Royal Statistical Society, Series B, 54*, 221-228.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g- and h-distributions. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Exploring data tables, trends, and shapes* (pp. 461-513). New York: Wiley.

Hogg, R. V., Fisher, D. M., & Randles, R. H. (1975). A two-sample adaptive distribution free test. *Journal of the American Statistical Association*, *70*, 656-661.

Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika*, *67*, 85-92.

Johnson, N. J. (1978). Modified t tests and confidence intervals for asymmetrical populations. *Journal of the American Statistical Association*, *73*, 536-544.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of Educational Researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research*, *68(3)*, 350-386.

Keselman, H. J., Wilcox, R. R., & Lix, L. M. (2003). A robust approach to hypothesis tesing. *Psychophysiology*, *40*, 586-596.

Keselman, H. J., Wilcox, R. R., Othman, A. R., & Fradette, K. (2002). Trimming, transforming statistics, and bootstrapping: Circumventing the biasing effects of heteroscedasticity and nonnormality. *Journal of Modern Applied Statistical Methods*, *2*, 288-309.

Liu, R. Y., & Singh, K. (1997). Notions of limiting P values based on data depth and bootstrap. *Journal of the American Statistical Association*, *92*, 266-277.

Lix, L. M., & Keselman, H. J. (1995). Approximate degrees of freedom tests: A unified perspective on testing for mean equality. *Psychological Bulletin*, *117*, 547-560.

Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and non-normality. *Educational and Psychological Measurement*, *58*, 409-429 (58, 853).

Luh, W., & Guo, J. (1999). A powerful transformation trimmed mean method for one-way fixed effects ANOVA model under non-normality and inequality of variances. *British Journal of Mathematical and Statistical Psychology*, *52*, 303-320.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156-166.

Othman, A. R., Keselman, H. J., Padmanabhan, A. R., Wilcox, R. R., & Fradette, K. (in press). Comparing measures of the "typical" score across treatment groups. *British Journal of Mathematical and Statistical Psychology*.

Othman, A. R., Keselman, H. J., Wilcox, R. R., Padmanabhan, A. R., & Fradette, K. (2002). A description and illustration of a test of symmetry. *Journal of Modern Applied Statistical Methods*, *2*, 310-315.

Rocke, D. M., Downs, G. W., & Rocke, A. J. (1982). Are robust estimators really necessary? *Technometrics*, *24(2)*, 95-101.

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers leverage points. *Journal of the American Statistical Association*, *85*, 633-639.

SAS Institute Inc. (1989). *SAS/IML software: Usage and reference, version 6* (1st ed.). Cary, NC: Author.

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error probabilities of the t test to departures from population normality. *Psychological Bulletin*, *111*, 352-360.

Welch B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*, 350-362.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*, 330-336.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing*. New York: Wiley.

Wilcox, R. R. (1988). A new alternative to the ANOVA F and new results on James's second-order method. *British Journal of Mathematical and Statistical Psychology*, *41*, 109-117.

Wilcox, R. R. (1994). A one-way random effects model for trimmed means. *Psychometrika*, *59*, 289-306.

Wilcox, R. R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? *Review of Educational Research*, *65(1)*, 51-77.

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego: Academic Press.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego: Academic Press.

Wilcox, R. R., & Keselman, H. J. (2003). Repeated measures one-way ANOVA based on a modified one-step M-estimator. *British Journal of Mathematical and Statistical Psychology*, *56*, 15-25.